模块一·知识与推理的根基

启发式、偏差与理性

你的心智在走捷径。一场持续五十年的争论由此开始：这些捷径是漏洞，还是高明之举？

琳达今年三十一岁，单身，心直口快，非常聪明。她主修哲学，学生时代深切关心歧视与社会正义，还参加过反核示威。现在，快速回答：哪一项更可能？A. 琳达是一名银行出纳员。B. 琳达是一名银行出纳员，并且积极参与女权运动。

大多数人会偏向 B。它很搭，像一段完整的故事。一九八三年，阿莫斯·特沃斯基与丹尼尔·卡尼曼抛出这道题时，超过八成的受访者选了 B，其中包括正式修过概率论的决策科学博士生。但再看一眼：每一位女权主义银行出纳员都「是」银行出纳员；B 组完全落在 A 组之内。合取事件的概率绝不可能高于它的任一组成部分。选 B 不只是错了，它违背了一条你明明认同的逻辑法则。你或许刚刚眼睁睁看着自己的心智违反了一条它自己也同意的规则。这道裂缝有个名字，叫做合取谬误[1]；而撬开它，正是今天的任务。

- 核心认知偏差
- 资源理性综合有前景
- 意志力燃料模型已崩溃

我们身在何处

前十日，我们搭好了良好推理的骨架：什么算作知道（第 1 日）、科学如何过滤主张（第 2 日）、三种推理引擎（第 3 日）、贝叶斯信念更新律与基础率陷阱（第 4 日），以及为什么每个模型都是一次有用的谎言（第 10 日）。这一切都属于「规范性」：心智应当如何推理。今天，我们把显微镜对准实际的工具。人类能达到这个标准吗？而当我们失足时，那是该修补的缺陷，还是心智在做比逻辑更聪明的事时留下的指纹？

纲领

我们思考时走的捷径

一九七四年，两位以色列心理学家在《科学》上发表了一篇论文，悄悄改写了一整个领域。特沃斯基与卡尼曼提出，人并不是靠「做概率题」来估计概率的；我们会顺手抓起一小把启发式[2]：通常奏效、偶尔以可预测方式失灵的心智捷径。其中三把承担了大部分重担。

代表性启发式[3]。我们凭一件事与心中原型的相似程度来判断它有多可能。琳达正

[1] 合取谬误是指把一个组合判断成比它所包含的某个部分更可能，尽管组合事件只是子集。
[2] 启发式是一种快速的经验法则，通过使用单一线索或模式来减少认知负荷，而不是做完整计算。
[3] 代表性启发式是按某个对象有多像心中原型来估计概率的捷径。

是这样被绊倒的：她太符合女权主义者的刻板印象，以至于「女权主义银行出纳员」听起来比「银行出纳员」更贴切；这种「合辙」的感觉悄悄压过了集合的算术。代表性启发式也会让我们忽视基础率，这正是第 4 日里的陷阱：听说「史蒂夫性格温顺、整洁、喜欢秩序」，人们猜他是图书管理员而非农民，却忘了农民的人数本来就远超图书管理员。

易得性启发式[4]。我们凭例子浮上心头的容易程度来判断一件事有多常见。空难登上新闻后，坐飞机感觉致命，尽管事故之所以成为新闻，恰恰说明它多么罕见。生动、新近、情绪化的事件会被过度加权，因为它们容易被想起。

锚定与调整[5]。估计一个未知数字时，我们会抓住手边的任何数值，再以此为起点调整，而调整往往远远不够。特沃斯基与卡尼曼转动一个被动过手脚的幸运轮，让它停在 10 或 65，然后问人们非洲国家占联合国的百分比。看到 65 的人猜得远高于看到 10 的人。那个轮子显然是随机的，与非洲毫无关系，却仍然锚定了他们。

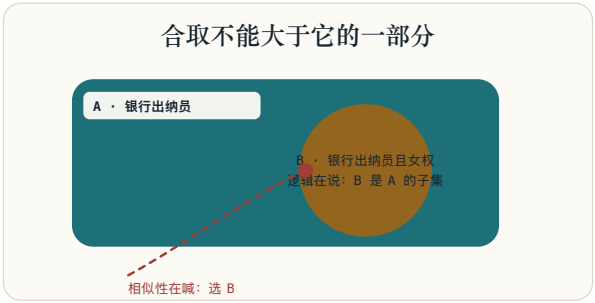
真正令人不安的不是我们会犯错，而是这些错误和视错觉一样顽固。在缪勒–莱尔错觉中，两条等长的线因为两端箭头方向不同而看起来不等长，而知道把戏并不能让它们看起来相等。卡尼曼的论点是，许多认知偏差就像这样：不是无知，而是你无法随意关掉的机器自动输出的结果。理解谬误并不会消除那种感觉。二〇〇二年，卡尼曼凭借这项工作获得诺贝尔经济学奖；一九九六年去世的特沃斯基若还在，想必会共享殊荣，但诺贝尔奖不追授。

静态替代·琳达机

交互面板让读者在「相似性镜头」与「集合逻辑镜头」之间切换。静态版本展示核心结论：生动的描述让较小的集合感觉更匹配，但「银行出纳员且女权主义活动家」始终是「银行出纳员」的子集。

[4] 易得性启发式是按例子多容易被想起来判断事件有多常见的捷径。

[5] 锚定与调整是从一个初始数值出发估计未知量，但后续调整通常不够充分。



琳达描述让 B 更有故事感，但集合关系没有改变：P(A 且 B) 永远不能超过 P(A)。

镜头	它看到什么	结论
相似性	琳达更符合女权活动家的故事，而不是普通银行出纳员的故事。	B 感觉更可信。
集合逻辑	每位女权主义银行出纳员都已位于银行出纳员集合之内。	B 不可能比 A 更可能。

模型及其裂缝

快与慢：谨慎使用

为了整理这一切，卡尼曼在他二〇一一年的畅销书《思考，快与慢》中推广了一出双主角戏剧。「系统 1」快速、自动、毫不费力、直觉式：它读出路牌上的字，察觉声音里的敌意，在你还没反应过来就冲琳达喊出「女权主义者」。「系统 2」缓慢、审慎、费力：是你填税表、或检查 B 是否真是 A 的子集时动用的那部分。按这种说法，偏差发生在忙碌的系统 2 没能审核快速答案的时候。

它是个绝妙的教学装置。但按字面理解，它很可能是错的，领域自己也知道。有三条裂缝值得命名，因为优秀的科学家会保留梯子，同时踢开脚手架。

也许根本不存在「两个」东西。「两个系统」从来都更像是隐喻，而非解剖结构。即便是主要支持者也有所退却：乔纳森·埃文斯与基思·斯塔诺维奇在二〇一三年一篇审慎的论文中，放弃了两个「系统」的说法，转而主张两种「加工类型」——一个更弱、更模糊的断言。二〇一八年，大卫·梅尔尼科夫与约翰·巴奇更进一步，称双过程类型学是「方便而诱人的迷思」，「缺乏经验支持」，并「系统地阻碍科学进步」。那个整齐的二分——意识/无意识、费力/自动——并没有聚成两捆清晰的神经束。

心智燃料箱是一场海市蜃楼。一个著名的配套观念，自我耗竭[6]，主张费力的自控依赖一种有限能量储备：现在忍住不吃饼干，下一次诱惑来临时你就会投降，因

[6] 自我耗竭指自我控制会消耗有限的意志力资源，从而使后续的自我控制变得更困难。

为你的意志力肌肉已经疲惫。它催生了上千项研究。然后，二〇一六年一项跨二十三个实验室的研究预登记复现发现，效应几乎为零：标准化效应量约 0.04，统计上与零没有区别。这是第 2 日的复现危机延伸进了今日话题。支撑「费力系统 2」的一块招牌发现，在严格检验下蒸发了。

所以，不必把这个模型攥得太紧。快与慢的区分仍是有用的简写，我们也会继续使用它，但要像第 10 日教我们的那样对待任何模型：它是有损地图，不是领土。更深层的动作已转移到别处。

争论

理性大论战

到这里，领域分裂成两个阵营，四十年间既有成果，也互相戳刺。

一边是卡尼曼与特沃斯基的「启发式与偏差」传统。它的信息大致是：人类直觉布满了系统性错误。认知偏差是认知错觉，是相对于逻辑与概率金标准的真实、可测量的失败。其实践取向是改良主义：既然心智会误导我们，就该建立纠正机制——培训、清单、专家系统，以及后来的助推[7]。

另一边是格尔德·吉仁泽与柏林 ABC 研究小组，他们回应：别急。他们主张，所谓「错误」往往并不是错误，而实验常常——虽非故意——对人类不利。他们有两步不同的论证，值得分开看。

第一步：换个格式，谬误就淡了

吉仁泽的第一个论点关于「信息格式」，直接呼应第 7 日。同样的内容，依编码方式不同，可携带的可用结构也大相径庭。以那些让人类——也让医生——显得很糟糕的基础率问题为例。用概率格式表达：

> 某种疾病影响 *1%* 的女性。一项检测在患者有病时 *90%* 能检出，但也会在 *9%* 的健康者身上出现假阳性。一位女性检测结果为阳性。她实际患病的可能性有多大？

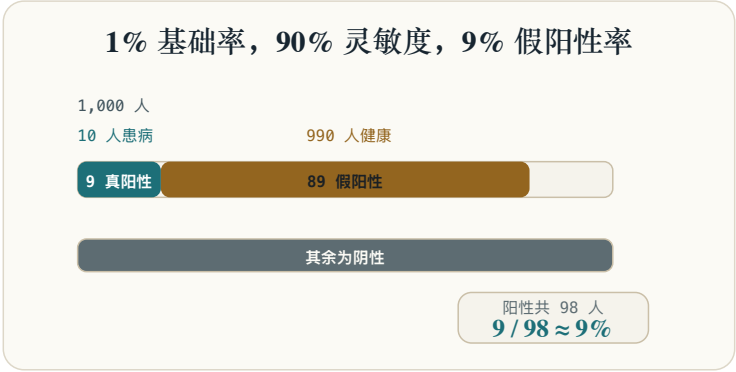
大多数人的回答落在 80%–90% 之间，许多医生在一次又一次的研究中也同样如此。正确答案约为 9%。但请看，当你把同样的事实从抽象百分比换成自然频率[8]：在 1,000 名女性中，10 人患病，其中 9 人检测阳性；990 名健康女性中，约 89 人也检测阳性。因此，在约 98 名阳性者中，只有 9 人真的患病。突然间，答案——9/98，约十一分之一——几乎就在眼前。你能数出来。二〇一七年的元分析发现，改用自然频率后，正确贝叶斯答案的比例从约 4% 提高到约 24%，大约翻了五倍。吉仁泽的要点是，基础率谬误并不只是人类线路里的固定缺陷。它部分是格式造成的假象：我们把心智不擅长消化的单次事件概率喂给了它。换一种表征，许多「非理性」就会消散。

[7] 助推是通过调整选择呈现方式来影响决定，同时不禁止任何选项的行为经济学干预。

[8] 自然频率把概率信息表达为参考类别中的计数，例如 98 次阳性检测中有 9 次为真阳性。

静态替代·自然频率网格

交互面板可改变疾病基础率，并在概率格式与自然频率格式之间切换。静态图固定经典 1% 场景，让贝叶斯答案可以直接数出来。



自然频率把贝叶斯答案变成计数题：在阳性检测者中，真正患病者只是所有阳性中的一小部分。

组别	每 1,000 人中的计数	阳性检测
患病女性	10	9 真阳性，1 漏检
健康女性	990	89 假阳性，901 真阴性
所有阳性检测	98	仅 9 人为真病例，因此后验概率约为 9%。

第二步：少即是多

吉仁泽第二个、也更大胆的论点是：简单启发式不只是可原谅的，有时还比统计学家偏爱的复杂、贪信息的方法更好。旗帜是快速节俭启发式[9]：忽略大部分可用信息，仍能做出出色决策的小规则。

代表作是凝视启发式[10]。外野手如何接住高飞球？不是靠测量速度、计算抛物线、再冲刺到落点。风与旋转会让实时计算不可行。相反，外野手盯住球，跑动时保持凝视角度不变。遵循这一条规则，你就会到达球下落的位置，无需微积分。这个启发式奏效，不是「尽管」忽略了信息，而是「因为」忽略了信息：它抛掉一切，只留下那条唯一关键的线索。

[9]快速节俭启发式是故意忽略大部分可用信息、只依赖少数有用线索的简单决策规则。
[10]凝视启发式是通过保持视线角度不变来追踪和拦截移动目标的简单规则。

再如识别启发式[11]：被问到两座城市哪座更大，如果你只认识其中一座，就押注你认识的那座。粗糙得可笑，然而因为「是否认识」往往与城市大小相关，它能击败复杂模型。在一个著名发现中，这甚至产生了「少即是多效应」：认识城市较少的人有时得分更高，因为他们能使用这条启发式，而那些什么城市都认识的人反而要依赖更不可靠的知识。

为什么忽略信息反而有帮助？深层答案由吉仁泽与亨利·布莱顿在二〇〇九年一篇题为 *Homo Heuristicus* 的论文中阐明，那就是偏差-方差权衡[12]——我们将在第 136 日的机器学习部分再次遇见它。参数众多的复杂模型对训练数据拟合得极好，但也拟合了噪声，所以在新数据上摇摇晃晃。简单启发式几乎什么都不拟合，因此也不会拟合噪声。它可能有偏差，但它稳定；而在嘈杂、小样本的世界里，稳定常常获胜。

重新框定

西蒙的剪刀

整场争论底下，藏着一个关于「标尺」的问题。卡尼曼说人不理性，意思是：用逻辑与概率法则来衡量。吉仁泽说人很聪明，意思是：用他们在真实环境中的成功来衡量。这是两把不同的尺子，而战争的大部分内容，正是在争哪把尺子合法。

最早看清这一点的是赫伯特·西蒙：经济学家、认知科学家、人工智能先驱，一九七八年诺贝尔奖得主。整个五〇年代，西蒙都在论证那个幻想——一个能优化所有选项的完全理性行动者——对任何真实生物来说都根本行不通。真实的心智受限于时间、记忆与注意力。所以我们不是「优化」，而是满意化[13]：找到一个够好的选项就停下来。西蒙称之为有限理性[14]，并给了一个悄然平息争论的意象：

> 「人类理性行为是由一把剪刀塑造的，剪刀的两片刀刃分别是任务环境的结构与行动者的计算能力。」 西蒙，1990。

一把剪刀只有在两片刀刃同时咬合时才能剪断东西。你不能单独评判其中一片。卡尼曼与特沃斯基主要研究了心智这片刀刃，并拿逻辑标尺衡量，发现它有欠缺。吉仁泽坚持必须同时看两片刀刃：心智与它正在裁剪的世界。一旦如此，昨天的「偏差」常常显露出它其实是与环境匹配的工具。两个阵营都不全错。他们只是在检查同一把剪刀的不同刀刃。

静态替代·西蒙的剪刀

交互图可在「只用心智刀刃评判启发式」与「用心智加环境共同评判」之间切换。静态图让两片刀刃都保持可见：行为由它们的接触产生。

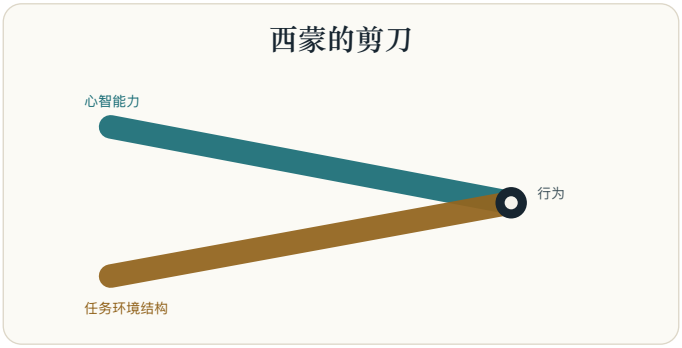
---

[11] 识别启发式是当两个对象中只认识一个时，用「认识」本身作为判断线索。

[12] 偏差-方差权衡指简单模型稳定地朝一个方向偏误，与复杂模型拟合噪声、在新数据上波动过大之间的张力。

[13] 满意化是指搜索到一个足够好的选项后就停止，而不是继续寻找理论上的最优解。

[14] 有限理性研究在时间、记忆、注意力、信息与计算约束下的理性行动。



只看一片刀刃：规则显得粗糙
看两片刀刃：规则可能正好适配环境

有限理性不是只研究心智缺陷；它研究有限心智和结构化环境如何一起产生可用行为。

评判标准	启发式看起来如何	示例结论
仅凭逻辑	它忽略了大部分信息。	非理性：规则太粗糙。
心智加环境	它利用了任务世界的结构。	聪明：凝视启发式以低廉成本解决了实时接球问题。

前沿·2026

三条活边，带上防炒作滤镜

这门课每一天都以前沿研究收尾，每条主张都标好了它能承受多少重量。理性之争并没有以条约告终；它成熟为更尖锐的经验问题。

边缘 01 ● 稳健偏差已确立 ● 意志力燃料模型已崩溃

哪些偏差挺过了复现危机？

启发式与偏差文献也没有逃过第 2 日的清算。但残骸呈现出一种意味深长的模式：像知觉式的一次性认知错觉保留了；脆弱、多环节的社会效应常常没有。

发现	证据状态	留存了什么
合取谬误	已确立	琳达效应在数十年、不同文化和多种表述中稳定复现，大型语言模型也会中招。
锚定	已确立	心理学中较稳健效应之一；现已获得计算层面解释。
自我耗竭	已崩溃	23 实验室研究预登记复现发现 d 约 0.04，实际无效应。
许多社会启动效应	有争议	部分概念启动结果未能复现；卡尼曼曾公开警告该领域。

教训不是「偏差不存在」。而是稳健的偏差往往是有机制的偏差，这也为接下来最有趣的发展搭好了舞台。

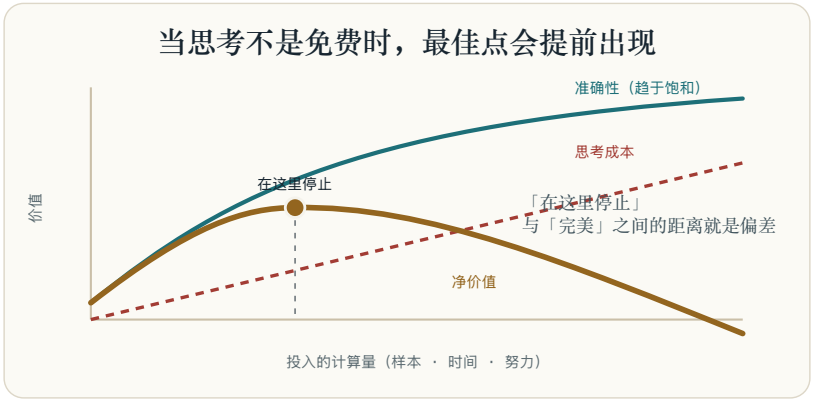
边缘 02 ● 资源理性综合有前景 ● 普适解释有争议

和平条约：偏差即最优偷懒的心智

这个领域近年来最重要的想法，是真想化解战争、而非赢得战争，而且它正落在我们计算线索的核心地带。它叫资源理性分析[15]，由福尔克·利德与汤姆·格里菲思二〇二〇年在《行为与脑科学》的目标文章提出。认真对待西蒙的洞见，并把它数学化：假设心智在追求准确，但思考要耗费资源——时间、记忆、计算。然后问，给定预算，一个行动者最优会使用什么策略。许多经典偏差开始看起来像廉价近似：用一点准确性交换大量速度。

最干净的工作示例是锚定。假设心智估计一个不确定量时，像统计学家一样：从概率分布中抽样，再取平均。但每抽一次都要耗费努力，所以最优行动者只抽少数几次就停。如果它从附近某个值——一个锚——开始抽样，并在样本还没走远之前就停下，估计值就会留在锚附近。这就是调整不足：锚定偏差，被推导成一个珍惜自己时间的心智的正确行为。缺陷，就是当思考不免费时，「最优」可能长成的样子。

[15] 资源理性分析把准确性与时间、记忆、计算等成本放在一起权衡。



资源理性分析把偏差解释为有限计算下的折中：多想一点会更准确，但不一定更值得。

这能承载多少重量？该框架有前景，且越来越丰产。它重新推导了锚定、某些记忆效应，甚至对「心智为何看起来像有两个系统」给出了理性重释。但请保持防炒作滤镜。因为资源理性有自由参数——成本、先验、算法——批评者警告，它可能变成不可证伪的：事后总能找到一套参数把任何行为合理化。那是第 2 日划界问题换了新衣：能解释一切的理论，什么也不能解释。

边缘 03 ● 大型语言模型镜像部分人类偏差 ● 相同推理解读有炒作风险

机器继承了我们的捷径

这是一个会让西蒙高兴的转折。当研究者把经典认知心理学任务搬到大型语言模型上时，模型的表现惊人地像我们。二〇二三年《美国科学院院刊》（PNAS）的一篇论文中，马塞尔·宾茨与埃里克·舒尔茨把琳达问题喂给 GPT-3，它犯了合取谬误，也表现出锚定与框架效应。后续研究发现，能力更强的模型犯这类直觉错误更少，但偏差并未消失。一个被训练来预测海量人类文本中下一个词的系统，显然不只吸收了我们的知识，也吸收了我们的部分推理反射。

抵制过度解读。大型语言模型复现合取谬误，并不证明它像人类一样推理。这些系统究竟进行了真正的推理，还是只是复杂的模式补全，正是留给第 139 日的争论。目前：这种镜像真实、可复现、引人入胜；如何解读则完全开放。它还给 AI 模块——第 138–145 日——抛出一个尖锐问题：当一台机器给出流畅、自信、错误的答案时，那是机器的毛病，还是一面照向我们的镜子？

未决问题

真正悬而未决的事

对心智实际如何推理的研究已经五十年，诚实的账簿仍然很长：

- 「是否存在唯一正确的理性标准?」抑或「理性」总是相对于目标与环境?
- 「一个过程还是多个过程?」心智是单一引擎在预算下做近似推断，还是确实存在不同模式？两个系统的图景受了伤，但尚无公认的替代者胜出。
- 「偏差该纠正还是该尊重?」如果偏差是有限心智的最优输出，去偏可能让你在现实世界中更糟，而在逻辑测试中更好。
- 「资源理性是解释，还是只是重新描述?」它能变得可证伪吗，还是总会找到一种成本函数，为人们的所作所为开脱?
- 「而 AI 模块继承的问题：」当一个系统复现我们的谬误时，它学到的是我们的推理，还是我们的残渣?

三句话总结今日

大观念

启发式既是有用捷径，也是可预测的陷阱；标尺决定我们讲哪种故事。

最佳类比

西蒙的剪刀：心智与环境共同剪裁理性。

当下争议

资源理性分析试图调和偏差研究与生态理性。
计算·演化·信息格式·失败的意志力能量模型·轻涌现。

明日 → 第 12 日

网络

今天研究了一颗心智及其所适应的环境。明天我们把镜头拉远，看向心智与事物之间的连接：六度分隔、枢纽与幂律、观念与流行病如何传播，以及真实世界网络是否真正无标度的争议。

来源与延伸阅读

来源

1. Tversky, A. & Kahneman, D. (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* 185(4157): 1124-1131.
2. Tversky, A. & Kahneman, D. (1983). "Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review* 90(4): 293-315. doi:10.1037/0033-295X.90.4.293. doi.org/10.1037/0033-295X.90.4.293
3. Simon, H. A. (1955). "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics* 69(1): 99-118.
4. Simon, H. A. (1956). "Rational Choice and the Structure of the Environment." *Psychological Review* 63(2): 129-138.
5. Simon, H. A. (1990). "Invariants of Human Behavior." *Annual Review of Psychology* 41: 1-19.
6. Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux.
7. Evans, J. St. B. T. & Stanovich, K. E. (2013). "Dual-Process Theories of Higher Cognition: Advancing the Debate." *Perspectives on Psychological Science* 8(3): 223-241.
8. Melnikoff, D. E. & Bargh, J. A. (2018). "The Mythical Number Two." *Trends in Cognitive Sciences* 22(4): 280-293.
9. Hagger, M. S. et al. (2016). "A Multilab Preregistered Replication of the Ego-Depletion Effect." *Perspectives on Psychological Science* 11(4): 546-573.
10. Gigerenzer, G. & Hoffrage, U. (1995). "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats." *Psychological Review* 102(4): 684-704.
11. McDowell, M. & Jacobs, P. (2017). "Meta-analysis of the Effect of Natural Frequencies on Bayesian Reasoning." *Psychological Bulletin* 143(12): 1273-1312.
12. Gigerenzer, G. & Goldstein, D. G. (1996). "Reasoning the Fast and Frugal Way: Models of Bounded Rationality." *Psychological Review* 103(4): 650-669.
13. Gigerenzer, G. & Brighton, H. (2009). "Homo Heuristicus: Why Biased Minds Make Better Inferences." *Topics in Cognitive Science* 1(1): 107-143.
14. Lieder, F. & Griffiths, T. L. (2020). "Resource-rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources." *Behavioral and Brain Sciences* 43: e1. doi:10.1017/S0140525X1900061X. doi.org/10.1017/S0140525X1900061X
15. Lieder, F., Griffiths, T. L., Huys, Q. J. M. & Goodman, N. D. (2018). "The Anchoring Bias Reflects Rational Use of Cognitive Resources." *Psychonomic Bulletin & Review* 25(1): 322-349.
16. Hertwig, R. & Grune-Yanoff, T. (2017). "Nudging and Boosting: Steering or Empowering Good Decisions." *Perspectives on Psychological Science* 12(6): 973-986.
17. Mertens, S., Herberz, M., Hahnel, U. J. J. & Brosch, T. (2022). "The Effectiveness of Nudging: A Meta-analysis of Choice Architecture Interventions." *PNAS* 119(1): e2107346118.
18. Maier, M., Bartos, F., Stanley, T. D., Shanks, D. R., Harris, A. J. L. & Wagenmakers, E.-J. (2022). "No Evidence for Nudging After Adjusting for Publication Bias." *PNAS* 119(31): e2200300119. 另见 Szaszi et al. (2022), e2200732119. doi.org/10.1073/pnas.2200300119
19. Binz, M. & Schulz, E. (2023). "Using Cognitive Psychology to Understand GPT-3." *PNAS* 120(6): e2218523120. doi.org/10.1073/pnas.2218523120